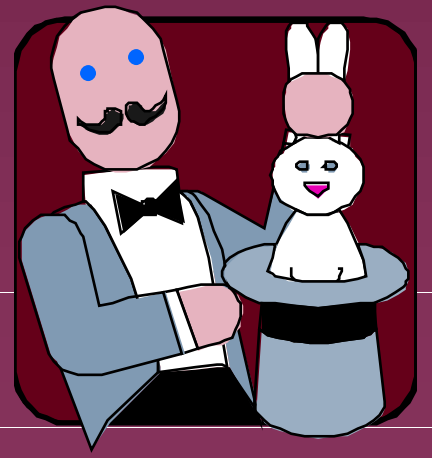
What do those weird XML types want, anyway?

25th International Conference on VLDBs,
Edinburgh, September, 1999

Steve DeRose, Chief Scientist
Brown Univ. Scholarly Technology Group;
Inso Corp. eBusiness Technologies
http://www.stg.brown.edu/~sjd



My perspective

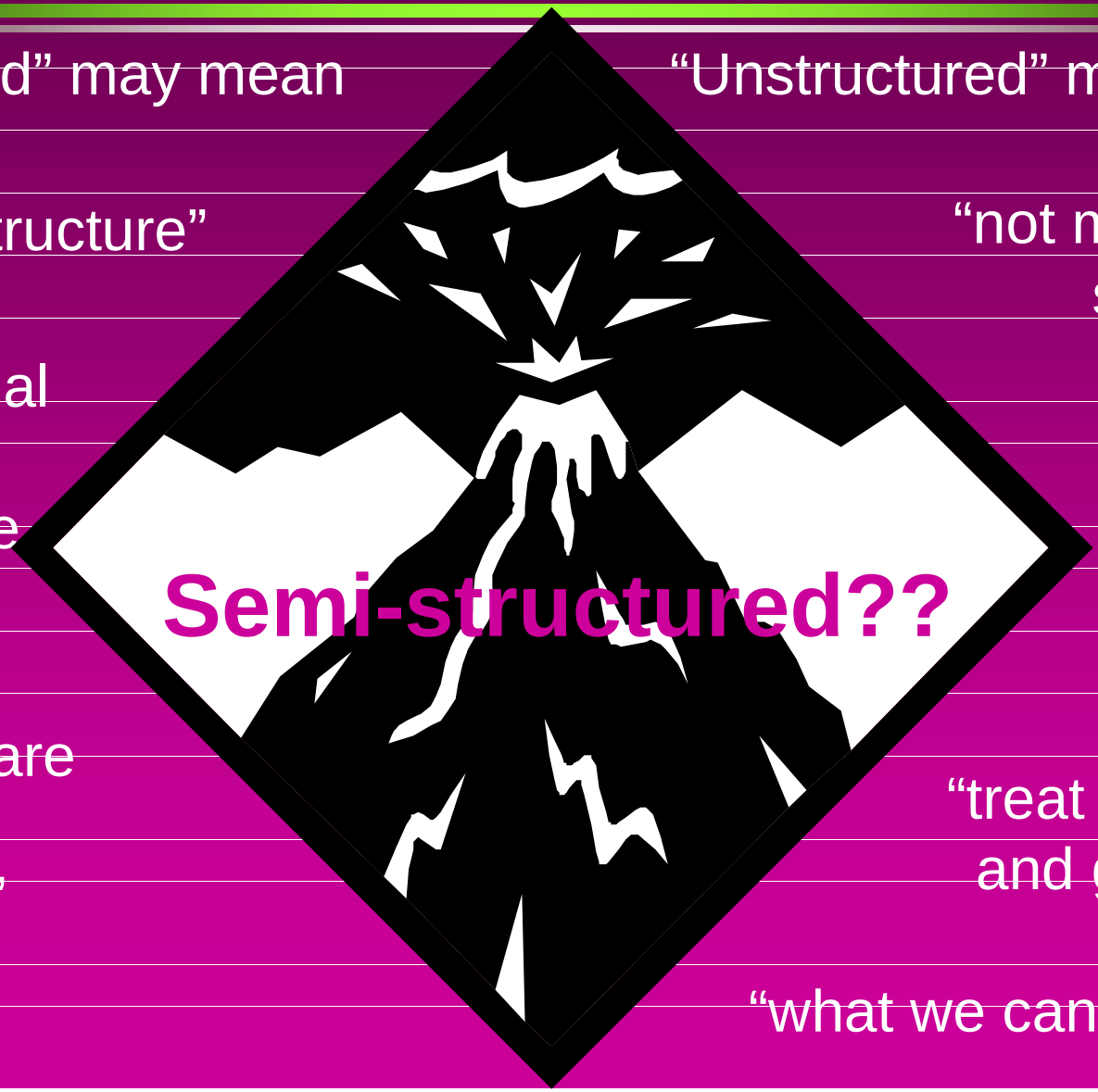
- Computer Science
  - Hypertext, FRESS
  - Formal languages, AI, NLP
- Linguistics
  - Corpus linguistics, stochastic models
  - Exegesis, translation theory and field linguistics
- Industry
  - Electronic Book Technologies, DynaText etc.
- Standards
  - DocBook, SGML, HyTime, EAD, SGML Open
  - XML, XPath, XPointer, XLink

Key points

- Documents not “structured” or “unstructured”
  - “Semi-structured” doesn’t quite do it
- Documents are natural language objects
  - What’s XML *about*, really?
  - What technical/practical issues?
- This won’t go away
  - What people produce, is what they will produce
  - 90% of corporate information
    - Surely even more in personal information
- Systems for the ambiguous/changing world
- Where to from here?

One view of the data world

- “Structured” may mean
  - “tabular structure”
  - “no optional or repeatable fields”
  - “all fields are quantities or enums”



“Unstructured” may mean

“not my kind of structure”

“too hard for me to analyze”

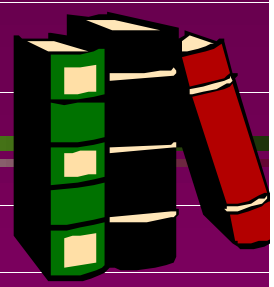
“treat as a blob and go home”

“what we can process”

Why just one axis?

- Kinds of structure
  - Tabular data
  - Networks/lattices/etc
  - Image/video data (raster/vector/model-based)
  - Long ordered series (time series, genomes)
- Not-so-familiar ones
  - Recursively partitioned series
    - (where *partitioning* is itself data; not just asn)
  - Link networks
    - Hypertext “nodes” have internal structure
  - Music (“unstructured”???)

What about documents?



- Everyday: books, catalogs, manuals, poems
- Documents don't fit the dichotomy
  - Is natural language "unstructured"?
    - (What's Chomsky been doing for 50 years?)
  - Is it "semi-structured"?
    - (note grimaces from document experts)
- Documents have a different *kind* of structure
  - How about "natural data"? (cf NLP, NT)
- XML is a representation for this type of structure

Bray: "Language is a text-based application"

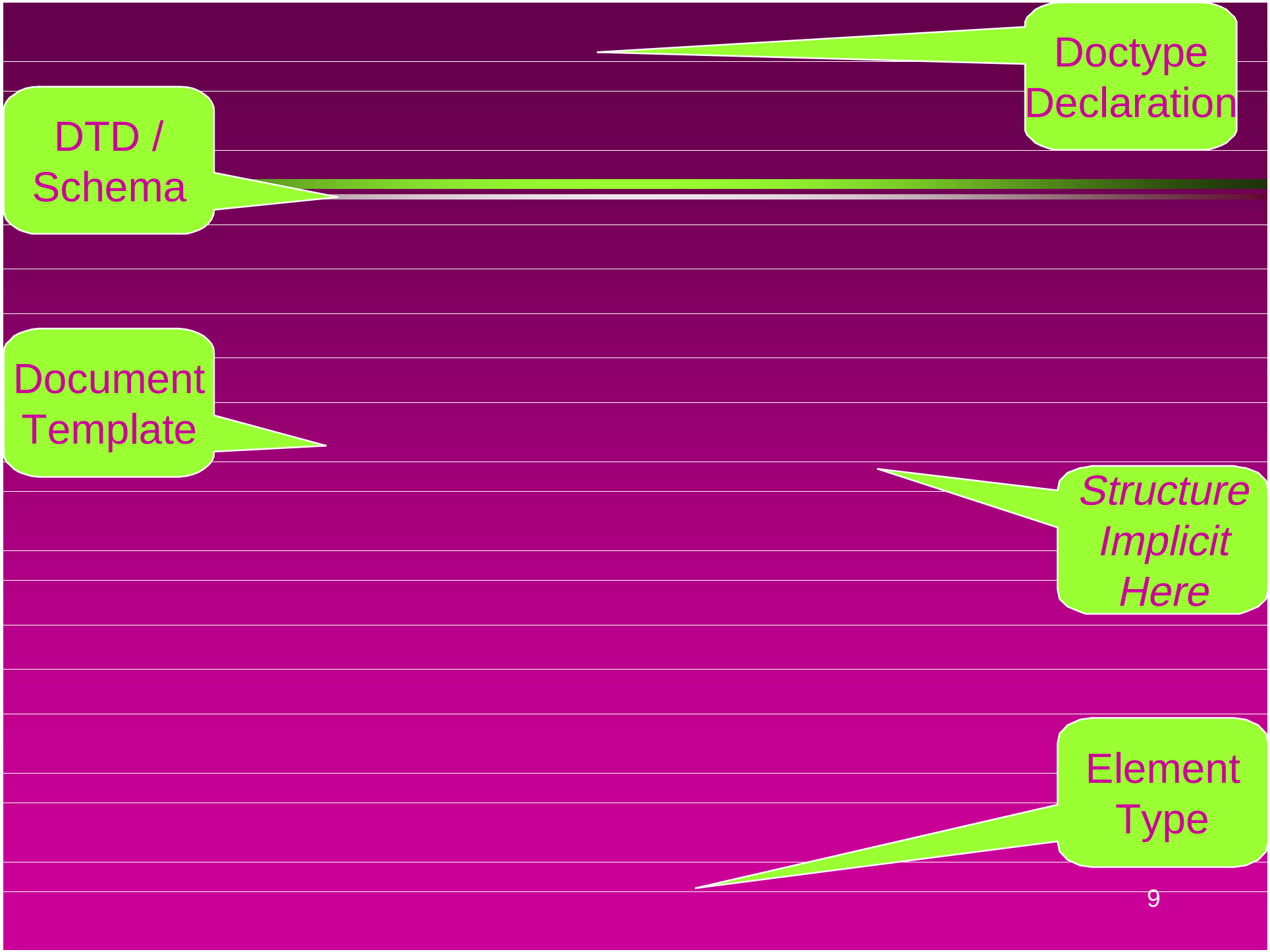
2-minute XML tutorial

- XML: Extensible Markup Language
- W3C Recommendation
- Standard way of labeling structural components of documents
- Mainly natural language objects
  - Manuals, books, novels, poems,…
  - Docbook, TEI, OEB
- But also
  - Transactions, KR, vector graphics, chem….

XML distinctives

Key characteristics	HTML?
Descriptive markup	Bipolar
Order matters	Yes
Hierarchical Structure	A little
Flexibility of tag set	NOT
Structure constraints	Only in theory
Human-readable	Usually

- Nutshell model
  - Interlinked recursive partitions
  - Natural-language objects
  - Yes syntax, but care about DOM


Doctype Declaration
DTD / Schema
Document Template
Structure Implicit Here
Element Type

Articles of XML faith

- Structural should reflect cognitive essentials
  - Formatting is important but an epiphenomenon
  - Structure therefore outlasts layout (one-way)
- Critical structural features
  - Order is a big deal
  - Hierarchies are useful (even cognitively real)
  - Links are useful, and related to structure
- And the odd one:
  - Structure ➡ extrinsic, unpredictable methods
    - (why XML can parse w/o schema)

XML model as a formalism

- **Document** = tuple (SystemID, PublicID, Element)
- **Node** = Element, Comment, PI, or Character
- **Element** = tuple (Name, Attributes, Content)
  - **Attributes** is unordered set of Attribute
  - **Attribute** = tuple of (Name, Value) where
    Name is unique within each *Attributes* (& declared)
  - **Content** = ordered list of Node
- **PI** = tuple (Target, Data)
- **Character** = as defined in Unicode

- Everything left is a String (list of Characters)

One subtlety: attributes

- Attributes
  - Unordered
  - Non-repeatable
  - Field-like
  - Scope: whole
  - Outside the partition
  - Properties-of
    - Has-property
  - Keys (IDREF)
  - *Element type is a special property*
- Elements
  - Ordered
  - Repeatable (95%!)
  - Language-like
  - Scope = extent
  - Exh. partitioned
  - Parts-of
    - Has-part
  - *Element itself just defined scope for attrs*

The usual suspects

- Examples I've seen given this week:
  - Few with repetition
  - One with ordering
  - One with attributes
  - None with text spanning leaves
  - None with multiple analytical perspectives
    - (two audience questions)
  - No indirect containment
- If XML examples come from RDBs,...
- "Image indexing" books may mean...

Better examples

- Literature
  - WWP
  - Philosophers
  - Biblical/Classical works
- Secondary materials
  - Dictionaries, corpora
  - E-Journals and E-Books
- Ephemera, letters, personal papers,…
- Linguistic annotation
- Multilingual texts

Where do we look?

- Suciu: cites ‘98, ‘99 on RDB repr’ns for XML
  - Each case in SGML/XML lit < ‘96
- Abiteboul: JLDB and parse vs. ast
- “Not much XML data” -- eh??
- Hierarchy papers vs. XML papers
- Kumar… & Gibson/Kleinberg/*Raghavan* HT98
- XML covers the range of un…semi…struc
- Humanities scholars have the interesting XML
  - *Computing and the Humanities* (MIT Press)
  - The Text Encoding Initiative (standard)
- Also a few companies, e.g. Boeing

The spherical cow

- Structure-aware querying/processing
  - Each type is different
- Graphics
  - String/quantity/boolean? Eh?
  - Keywords are a hack
- Long series
  - Order, interpolations, recurrence patterns
- What are the tempting spherical cows for documents?

Myths about XML

- “XML is a syntax that…”
  - XML people care about the model
- “XML is an interchange format”
  - Common use, but misses the point
- “XML is basically HTML cleaned up”
  - Philosophy/model is radically different
- “XML is aimed at rendering”
- “Order is a detail” (more later)
- “XML is just semi-structured data”
  - Not “no schema”, but a very specific class

What issues does XML raise?

- Theoretical issues
  - Topology
  - Boundaries
  - Discontinuity
  - Meanings
- Practical issues
  - What's "very large"
  - Implicit and pragmatic structure
  - Query formation
  - The datascape

T1: Document topology

- Order
  - Ubiquitous
- Hierarchy
- The recursive partition
  - (as data, not result)
- Text vs. numbers
- Multiple structures

T2: Boundaries

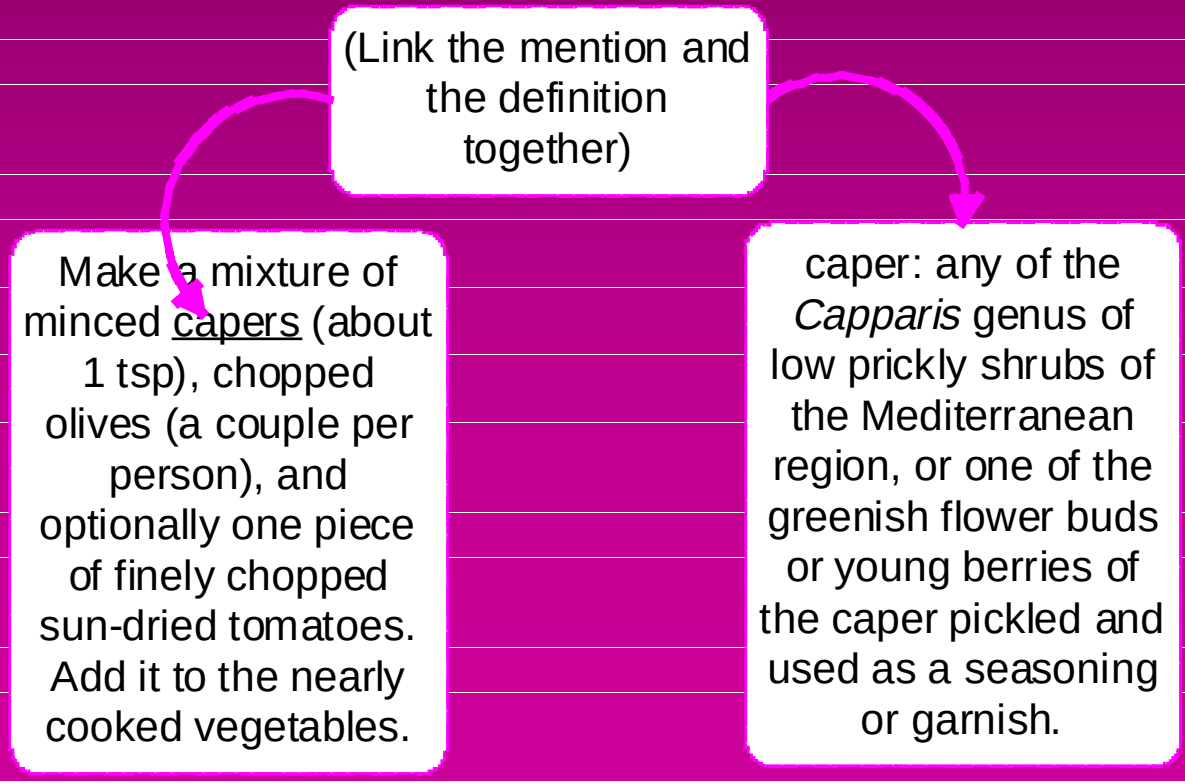
- Word - not as easy as it looks
- Markup units - which amount to word/sentence/etc bounds? *Unsolved*
  - God is now<fn>here we see…</fn>
  - God is now
here we see…
  - God is now<fn>…</fn>here we see…
- Cohesive units - proximity in big hierarchies
  - (big problem in XML search)
- Linguistic discourse unit analysis

T3: Discontinuity

- Footnotes, digressions, etc.

<s>Those are dog <fn>that is, 7-year</fn> years.</s>

- Hypertext
  - External links
- 3rd party annotation



T4: Complex meanings

- Polysemy
- Polyform
  - Many ways of expressing
  - “Sometimes” synonymy
- Pronouns, shortening, indirect reference
- Time/state dependent meanings
- Implicit information / pragmatics
- Multiple analytical perspectives
  - Book/chapter/verse vs. pericope/speech/sentence...

These are the characteristics of natural language

Even here…

- Very large data bases : proceedings / International Conference on Very Large Data Bases. 1977: Data base ; v. 9, no. 2; 1977: SIGMOD record; v. 9, no. 4. Notes: ... **Subtitle varies**
- Systems for large data bases: proceedings of the 2nd International Conference on Very Large Date [sic] Bases
- Very Large Data Bases: Proceedings International Conference on Very Large Data Bases
- Very Large Data Bases: 8th Intl Confernece on Very Large Data Bases Mexico City, Mexico
- Proceedings Vldb 83 Very Large Data Basis Conference Proceedings: Singapore 84 (Vldb-84)
- Very Large Data Bases: Proceedings, 11th International Conference on Very Large Data Bases
- Very Large Data Bases: Proceedings, 12th International Conference on Very Large Data Bases
- Proceedings of the Thirteenth International Conference on Very Large Data Bases, Brighton, England, 1987
- Proceedings of the Fourteenth International Conference on Very Large Data Bases
- Proceedings Vldb 89 International Conference on Very Large Data Bases

Even here, cont'd

- Very Large Data Bases: 16th International Conference on Very Large Data Bases/Proceedings: August 13-16, 1990, Brisbane, Australia
- Very Large Data Base Conference Proceedings 1991 (#Vl91)
- Proceedings of the Seventeenth International Conference on Very Large Data Bases: September 3-6, 1991: Barcelona (Catalonia, Spain)
- Very Large Data Bases, '92:
  Proceedings of the 18th International Conference on Very Large Data Bases, August 23-27, 1992 Vancouver, Canada
- Proceedings 19th International Conference on Very Large Data Bases: August 24Th-27th 1993, Dublin, Ireland
- Proceedings of the 20th International Conference on Very Large Data Bases: 20th Vldb Conference September 12-15, 1994 Santiago-Chile (#Vl94)
- Proceedings of the International Conferences on Very Large Databases Held in Zurich, Switzerland: Vldb-95
- Proceedings of the International Conferences on Very Large Databases Held in Bombay, India
- Proceedings of the Twenty-Fourth International Conference on Very Large Databases: New York, Ny, USA 24-27 August, 1998 (24th Conf)

Some problems with this

- Morphology and alternate forms
  "database" / "databases" / "data base" / "data bases"
  "11th" / "1985" / "85"
- Different representations of the "same" datum
  "24" / "24th" / "twenty-fourth"… / 1998
- Structural issues
  Date placement; multiple dates
- Missing or incomplete data
  editors, authors, "et al.", locations

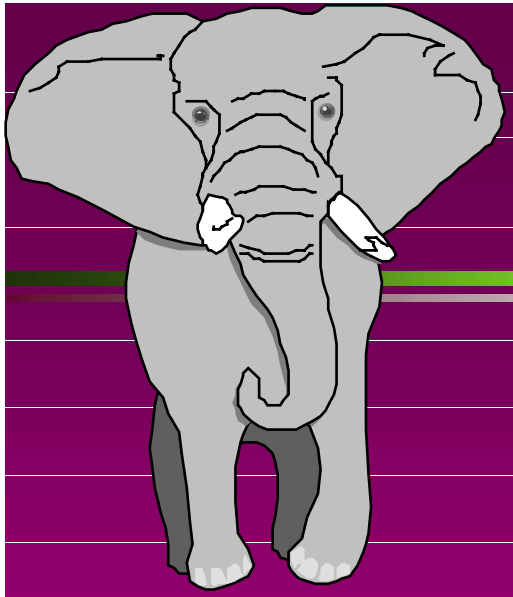
Ambiguity is information

Sir,

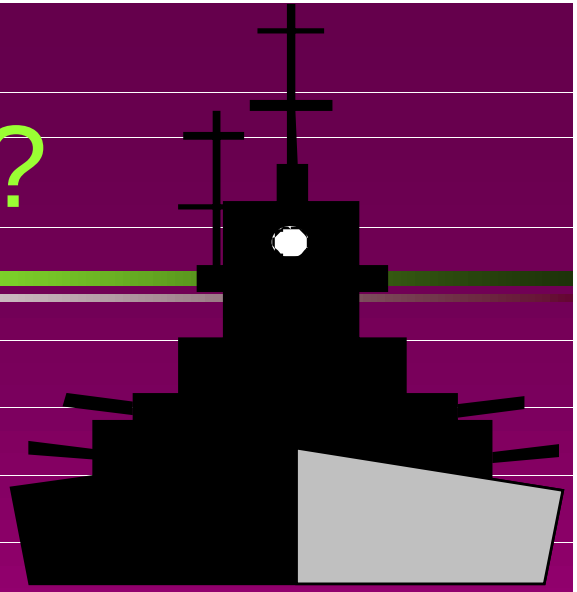
Mons. Compigne, a Savoyard by birth, a friar of the order of Saint Benedict, is the man who will present to you as his passport to your protection, this letter. He is one of the most discreet, the wisest and the least meddling persons that I have ever known or have had the pleasure to converse with. He has long earnestly solicited me to write to you in his favor, and to give him a suitable character, together with a letter of credence; which I have accordingly granted to his real merit, rather I must say, than to his importunity; for believe me, Sir, his modesty is only exceeded by his worth, I should be sorry that you should be wanting in serving him on account of being misinformed of his real character; I should be afflicted if you were as some other gentlemen have been, mislead on that score, who now esteem him.

Elizabeth went to Essex.
She had always liked Essex.

<persName>Elizabeth</persName> went to <placeName>Essex. </placeName> She had always liked <placeName>Essex.</placeName> <note type='uncertainty' resp='MSM'>It is not clear here whether <mentioned>Essex</mentioned> refers to the place or to the nobleman. -MSM</note>



P1: “very large”?

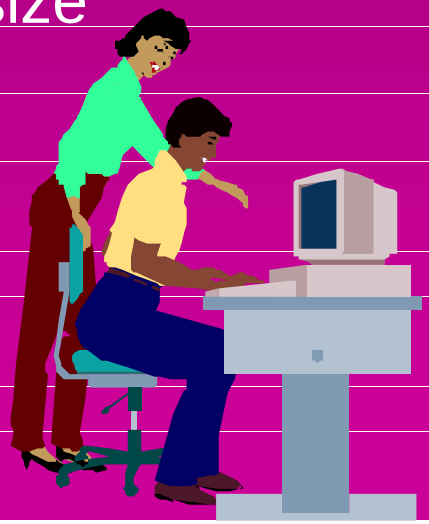


- The meaning of ‘very large’ is always changing
  - Brown Corpus vs. TLF, BNC, etc
  - Personal data production
- Individuals will have ‘very large’ information
  - Cost of having everything
  - Most of this info is documents
    - email, letters, family pictures,…
  - Tools now require planning and maintenance

P2: Implicit / pragmatic structure

- How do people speak?
  - Without consciously planning syntax
  - May not recognize/describe structure
  - Ambiguity is essential, not an error
  - Context
    - "Pragmatics"
    - No bound to relevant discourse size

- People author the same way
  - It won't change
  - XML authoring is not yet democratic



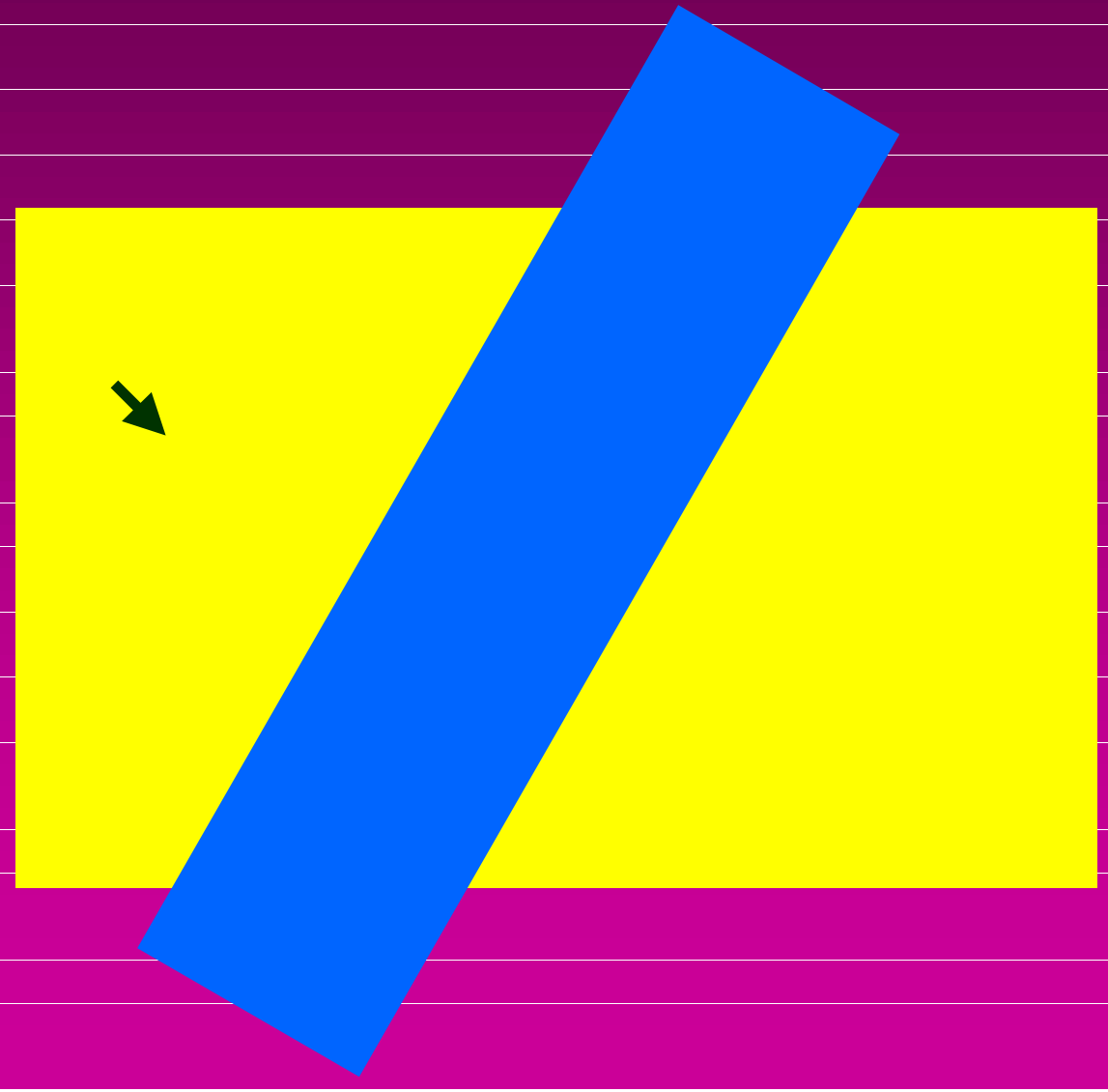
The writing process

- In a WP:
  - Think “new paragraph” etc.
  - Recall how to format it
  - Recall how to get that effect
  - Do action(s) to produce the effect
- In SGML/XML:
  - Think “new paragraph” etc. just as in WP
  - Recall or recognize element type
  - Apply that element type
- CACM 11/87 Coombs/Renear/DeRose

Why not ubiquitous?

- Cultural?
  - We've accommodated to bad design
  - We're used to dumb typewriters
- Cognitive?
  - Visual result seems direct (correlation)
- Freedom?
  - I want to produce weirdo effects at will (??)
- Playfulness, gadget fixation?
  - I get to figure out how to make it do this
- Instant gratification?
  - I see my result now; next year is next year

Reid’s example



Reid’s example


The line makes people think it would separate the two box pieces and make it so one would go away.
I don’t think so.

P3: Query formation

- People ask, as they speak
  - Syntax not the problem
  - Ambiguity
  - Polyform
  - Unconsciousness of own structure
- Interpretability depends on understanding
  - user and data context
  - implicit information
  - the data's relationship to all this

Some queries in XML

- “Find the parochial scholar”
  - Find each CHAPTER in this book, where all the FOOTNOTEs in the CHAPTER contain
    - AUTHOR elements with the same content.
    - AUTHOR elements with "formalname" attributes that point to the same AUTHOR_OF_RECORD element.
  - Issue
    - quantification over intermediate results

Queries, cont'd

- Find CHAPTERs where >75% of the footnoted authors are ones that are referenced from BOZO elements in my BOZOLIST document.
  - 3rd-party "kill files" for documents that only cite sets of authors you despise
- Same question but author's names appear in the headers of their (linked) documents.
- Issue:
  - linked vs. included data

Order queries

- Rhetoric of formal argument (TEI, RDF?)
  - <proof> cont (
    <step> cont <var> a “|” <var> b
    before <step> cont “~”a OR “~”b
- Literary ambiguity (TEI)
  - <placename> a before a as <persname>
    before all a linked from doc D
    by <link role="ambiguity">
- Organic chemistry (CML)
  - CH_2OH, | <bond> cont H and C | > 27, CH_2OH
- Bibliography (TEI, SGML-MARC, EAD,...)
  - Doc c cont <author> church, doc d cont <author>
    derose where c.pubdate = d.pubdate

One issue: binding value vs. leaf vs. subtree

XML query languages

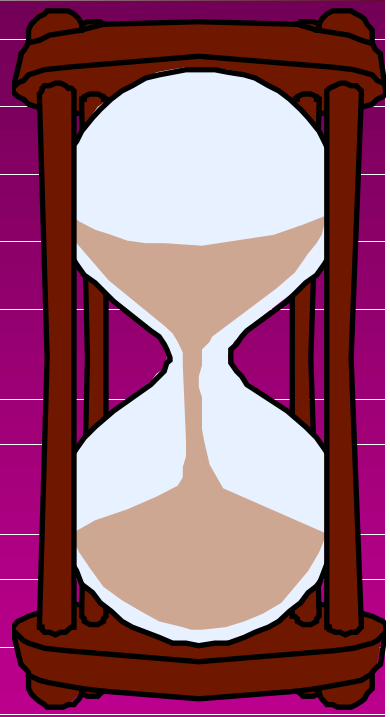
- Several mappings to RDB
  - Query depends on mapping
  - So what are we *actually* querying?
- XQL vs. Lorel
  - (I'll take Lorel)
- Avoid spherical cows

XPath/XPointer

- W3C expression lg for querying on structure
  - Primary input to new W3C query lg group
- Xpath (origins)...
  - Predicates over XML structure
  - Genetic navigation
  - Results are sets of locations (!= data)
- XPointer
  - Non-node (range) locations
  - Encapsulation into URLs
- `w3.org/TR/foo.xml#xptr(id('foo')/child(SEC)/…)`

P4: The datascape

- Data doesn’t stay the same
- Paper editions, manuscript variants,…
- Online:
  - Changing data, schemas, names, keys
  - Moving old data out of the way
  - Data you don’t realize you have
- Solutions
  - High-end: Attention over long spans
  - Low-end: ???



Preservation???

- People don’t back things up
- Software doesn’t help much
- Many almost-duplicates (implicit versions)
- Data you forget: Aliases, bookmarks, macros,...
- Binary data *formats* go away
- Media shelf lives

High-cost databases motivate data maintenance; but what of the rest of society?

Research items

- Theory
  - Recursive partition structures
    - Statistical algorithms that deal with scope
  - Discontinuities
    - Links, digresssions, annotations, footnotes
  - Persistent reference
  - Distributed querying
- Practice
  - Tools for data change and preservation
  - Tools for getting out of obsolete formats
  - Tools for versions (data, file, apps, file-system)

Summary

- Structural types
- Documents are linguistic objects
- Change and preservation
- Intersection
- Canaries

What XML world must do

- Publish/read in DB world
- Articulate the hard queries
- Provide the interesting data
- Push real XML, not HTML/tabs
- Be more formal
- Make friends in the DB world

What DB world can do

- Read/publish in the humanities computing world
  - (CHum www.oasis-open.org/cover)
- Take order and repetition very seriously
  - (using in examples would help)
- Don’t get stuck on interchange applications
  - (way too easy, even if loud)
- Querying XML, not RDB made *from* XML
  - (even if that’s underneath)
- Make friends in humanities computing
  - (ALLC/ACH 2000, Glasgow, July 21-25)


XML
VLDB

Some essential reading

- *Computing and the Humanities* (journal)
- André, Jacques, Richard Furuta, and Vincent Quint (eds). 1989. *Structured Documents*. Cambridge: Cambridge University Press. ISBN 0-521-36554-6.
- Coombs, James H., Allen H. Renear, Steven J. DeRose. 1987. “Markup Systems and the Future of Scholarly Text Processing.” *CACM* 30(11): 933ff.
- DeRose, Steven J., David G. Durand, Elli Mylonas, Allen H. Renear. “What is Text, Really?” *Journal of Computer Documentation*, Summer 1997. ACM.
- Gibson, David, Jon Kleinberg, Prabhakar Raghavan. 1998. "Inferring Web Communities from Link Topology."*Proc’s of Hypertext '98,* Pittsburgh. ACM..
- Ide, Nancy and Jean Veronis (eds.). 1995. *Text Encoding Initiative: Background and Context*. Boston: Kluwer Academic Publishers. 0792336895.
- Reid, Brian. 1981. *Scribe: A Document Specification Language and its Compiler*. Ph.D. thesis, Carnegie-Mellon University, Pittsburgh, PA. Also available as Technical Report CMU-CS-81-100.
- Tajima, Keishi, Yoshiaki Mizuuchi, Masatsugu Kitagawa, and Katsumi Tanaka. 1998. “Cut as a Querying Unit for WWW, Netnews, and E-mail.” In *Proc’s of Hypertext 98*. Pittsburgh. ACM Press.